Integrating Lexical and Prosodic Features for Automatic Paragraph Segmentation

Catherine Lai[a,*], Mireia Farrús[c], Johanna D. Moore[b]

[a]*Centre for Speech Technology Research, University of Edinburgh, Edinburgh, United Kingdom*
[b]*School of Informatics, University of Edinburgh, Edinburgh, United Kingdom*
[c]*TALN Research Group, DTIC, University Pompeu Fabra, Barcelona, Spain*

Abstract

Spoken documents, such as podcasts or lectures, are a growing presence in everyday life. Being able to automatically identify their discourse structure is an important step to understanding what a spoken document is about. Moreover, finer-grained units, such as paragraphs, are highly desirable for presenting and analyzing spoken content. However, little work has been done on discourse based speech segmentation below the level of broad topics. In order to examine how discourse transitions are cued in speech, we investigate automatic paragraph segmentation of TED talks using lexical and prosodic features. Experiments using Support Vector Machines, AdaBoost, and Neural Networks show that models using supra-sentential prosodic features and induced cue words perform better than those based on the type of lexical cohesion measures often used in broad topic segmentation. Moreover, combining a wide range of individually weak lexical and prosodic predictors improves performance, and modelling contextual information using recurrent neural networks outperforms other approaches by a large margin. Our best results come from using late fusion methods that integrate representations generated by separate lexical and prosodic models while allowing interactions between these features streams rather than treating them as independent information sources. Application to ASR outputs shows that adding prosodic features, particularly using late fusion, can significantly ameliorate decreases in performance due to transcription errors.

Keywords: Discourse structure, paragraph segmentation, prosody, spoken language understanding, coherence

1. Introduction

Audio and video recordings are increasingly popular ways to disseminate information. However, these sorts of spoken documents, such as podcasts or lectures, can often contain long passages. This can be difficult to browse and analyze without effective access to internal discourse structure. Previous work on automatically detecting this sort of discourse structure has generally focused on coarse-grained topic or story level segmentation (Tür et al., 2001; Tsunoo et al., 2017). Nevertheless, finer-grained multi-sentence segments are also important units of speech. For example, paragraph segmentation is valuable for automatic summarization (Sporleder and Lapata, 2006) and improving the readability of transcripts (Pappu and Stent, 2015). In fact, misplaced paragraph boundaries can increase the amount of effort it takes for human readers to identify the main points in texts, especially for unfamiliar subject areas (Goldman et al., 1995). Thus, an improved understanding of paragraph segmentation would also be beneficial for natural language generation. However, relatively little work has been done on automatic paragraph segmentation in text, let alone for speech.

In general, paragraph segmentation provides a good test case for teasing out how discourse structure is signalled using different aspects of speech. Qualitative analyses of text structure have found that paragraphs are internally cohesive units (Giora, 1986; Ji, 2008; McGee, 2014). However, this type of cohesion does not necessarily correspond to the notion of lexical cohesion (i.e., lexical similarity) used in coarse-grained topic segmentation. As such, previous work on automatic paragraph segmentation has mostly focused on predicting boundaries from surface features of individual sentences and their immediate neighbours. While this is effective for some genres, e.g., English news text, results can be highly variable depending on the domain (Sporleder and Lapata, 2006). So, we would like to obtain more general cues for supra-sentential cohesion and structure within paragraph segments. Prime candidates for this are

*Corresponding author
Email address: c.lai@ed.ac.uk (Catherine Lai)

discourse-oriented linguistic elements such as cue words, e.g., discourse markers such as *'because'* and *'otherwise'*, and prosody, i.e., pitch, loudness and timing aspects of spoken language (Passonneau and Litman, 1993b).

Previous work has shown that features based on cue words and prosody are helpful for segmentation tasks (Eisenstein and Barzilay, 2008; Hsueh et al., 2006; Tür et al., 2001, *interalia*). However, how they can be best utilized, together with other lexical and acoustic features, is still an open question. On the one hand, previous work on topic segmentation has argued that lexical and prosodic features contribute independent evidence and, thus, can be modelled separately (Tür et al., 2001). On the other hand, discourse analyses suggest that the prosodic form of a cue word can also be important for discourse interpretation (Hirschberg and Litman, 1994). On balance, we would expect these sorts of lexico-prosodic interactions to be important for paragraph structure. Moreover, previous analyses of paragraph structure also suggest that we need to track subtle topic related changes across potential boundaries (Giora, 1986). Thus, to investigate this, we need to be able to model interactions between features across time and across modalities.

In the following, we build on our previous work (Lai et al., 2016; Farrús et al., 2016) to investigate the predictiveness of cue words, supra-sentential prosody, and lexical coherence based features for automatic paragraph segmentation using a large corpus of TED talks.[1] In particular, we examine whether lexical and prosodic features that help high level topic segmentation are also predictive at the paragraph level. We frame paragraph segmentation as a binary classification task: for each utterance in a talk, we predict whether or not it is the last in a paragraph. Evaluation is performed with respect to human transcriptions of the talks.

TED talks are well known for being well-structured and entertaining to listen to. We consider these talks semi-spontaneous as they are prepared in advance and speakers are coached to be engaging and convincing in spoken form. However, talks still vary greatly in the style of their delivery. So, we expect features that are indicative of paragraph breaks in this dataset to be robust across a range of lecturing styles. The polished nature of these talks also makes it more likely that text-oriented discourse segmentation methods will work on this data set. This makes our work more comparable to previous work on topic and paragraph segmentation of text, and allows us to focus on studying the role of lexical and prosodic cues in this task.

[1] `http://www.ted.com`

The current work focuses on how we can integrate lexical and prosodic features effectively for this task. In particular, we test the hypothesis that discourse-oriented cues are better indicators of paragraph structure than traditional lexical similarity measures. We also expect discourse cue words and speech prosody to be more robust predictors of structure than the surface, syntax and language model based lexical features used by Sporleder and Lapata (2006) for text segmentation. Beyond this, we expect that modelling sequential information will improve performance and that allowing low level interactions between lexical and prosodic features will produce better results than modelling these information sources separately.

We also investigate whether prosodic features can help deal with error prone automatic transcriptions. Since our primary interest for this paper is establishing how we can make use of different types of speech features and model architectures, we focus on performance on sentence instances as determined by our human transcriptions for comparability across experiments,. However, we acknowledge this approach still leaves us quite some distance off a fully end-to-end automatic paragraph segmenter: this paragraph segmentation pipeline relies on the accuracy of transcription, punctuation restoration and the sentence segmentation based on that. As such, we consider our current work as providing necessary groundwork for developing an end-to-end discourse segmenter incorporating prosodic and lexical features, while also providing useful insights into how discourse structure is signalled in speech. Of course, automatic transcription and punctuation restorations are research problems in their own right. We leave a more detailed investigation of how completely automating upstream processes and joint modelling would affect our approach to future work.

In Section 2, we review previous work on paragraph segmentation and related work on speech and text segmentation. We describe the experimental setup used to test our hypotheses and research questions in Section 3. Section 4 describes results from our experiments using Support Vector Machines (SVMs), AdaBoost decision tree ensembles, Multi-Layer Perceptrons (MLPs), and Long Short-Term Memory recurrent neural networks (LSTMs). We also investigate different fusion strategies for lexical and prosodic features using LSTM based architectures (Section 4.4), and how well our automatic paragraph segmenters perform on automatic transcriptions (Section 4.5). We discuss the implications of our results and potential extensions of our work in Section 5. Section 6 concludes the paper.

2. Background

In order to determine the approach we should take to test our hypotheses, we review previous work on paragraph segmentation, as well as related work on the usefulness of lexical cohesion, prosody and cue word based features for related speech and text segmentation tasks. This section also reviews work on incorporating lexical and prosodic features for these tasks.[2]

2.1. Text Based Paragraph Segmentation

Previous work on automatically detecting paragraph structure has focused on text segmentation using supervised learning methods. In a comprehensive study, Sporleder and Lapata (2006) predict paragraph boundaries in texts from multiple genres (fiction, news, parliament) and languages (English, Greek, German). They employ BoostTexter (i.e., AdaBoost) based classifiers to investigate a range of surface form, syntactic, and language model based features. Overall, they find that surface features outperform syntactic features. However, a large amount of domain variation was observed: indicators for the first three words in a sentence provided the best performance for news and parliament data, while punctuation/quote marks appear more important for fiction. Nevertheless, their best results were obtained using the full feature set. This suggests that combining predictions from a number of individually weak predictors is important for this task.

Sporleder and Lapata (2006) show that good results can be obtained using relatively shallow, syntactically oriented, language processing methods. However, given the supra-sentential nature of the task, we might expect to see gains from incorporating more discourse-oriented language processing. In this vein, Filippova and Strube (2006) find that pronominalization/reference information and information structure (expressed syntactically in German) improves paragraph segmentation of German Wikipedia biographies using a similar BoostTexter set up. However, they find that features based on pre-determined lists of discourse connectives (e.g. *'finally'*, *'apart from'*, *'otherwise'*) produced worse performance than simply including the identity of the first few words of a sentence. This is inline with findings from Sporleder and Lapata (2006), who show that sentence initial phrases associated with paragraph breaks vary greatly between domains. So, it appears that the success of this approach depends heavily on whether domain specific cues are available, and performance drops on more heterogeneous data sets (e.g., fiction).

To capture more general lexical cues for paragraph breaks, it seems necessary to incorporate more contextual information. However, the features included in Sporleder and Lapata (2006) and Filippova and Strube (2006) focus on characteristics of the target sentence and its immediate neighbours. Furthermore, their classification method does not take advantage of the fact that sentences within a document come in a sequence. To address this, Shi et al. (2007) propose the use of discriminatively trained Semi-Markov Models, where paragraphs are predicted based on segment (edge) and boundary (node) features. In that work, boundary features represent observed sentence features similar to those used by Sporleder and Lapata (2006). In contrast, segment features depend on a hypothesized segmentation over the sentence sequence. In this case, paragraphs are characterized by their length, entropy, and the cosine similarity between a paragraph and its neighbours. This approach results in markedly better performance than SVM classifiers using boundary features alone, indicating that the sequence modelling approach is useful for this task. Shi et al. also report improved performance compared to Sporleder and Lapata (2006) for segmenting English novels, although performance is slightly worse for German novels. So, it seems that further investigation of features and models that can incorporate more contextual/sequence information is warranted.

2.2. Lexical Cohesion

The use of paragraph and sentence similarity features in Shi et al. (2007) assumes that paragraphs are *lexically cohesive*: sentences within a segment are lexically similar. These sorts of lexical similarity measures form the backbone of most text-based automatic topic segmentation methods. For example, in *TextTiling* (Hearst, 1994), topic boundaries are determined by identifying points of low lexical similarity between consecutive blocks of text. Similarly, the approach of Shi et al. (2007) is reminiscent of the divisive clustering algorithm used in *C99* (Choi, 2000), which attempts to maximize segment lexical similarity. In this vein, generative approaches have improved topic segmentation performance by fitting the observed data to models that associate each topic to a unique language model, and each segment to a topic (Allan et al., 1998; Yu et al., 2016) or mixture of topics (Purver et al., 2006; Georgescul et al., 2008).

While paragraphs are generally considered to form cohesive semantic units (Van Dijk, 1982), paragraph breaks rarely coincide with the types of segments usually consid-

[2]Note: In the automatic discourse *parsing* literature 'discourse segmentation' generally means clause level elementary discourse units, rather than the supra-sentential units we are interested in here.

ered in automatic topic segmentation (e.g., news stories, meeting agenda items). Analyzing English narratives, Ji (2008) shows that paragraph breaks often coincide with more subtle shifts related to time, location, and physical or mental states. Similarly, in an analysis of literary texts, Giora (1986) finds that paragraphs introduce new events rather than topics. In this vein, Sporleder and Lapata (2006) show that their supervised approach for paragraph segmentation performs better than applying an unsupervised lexical cohesion based topic segmenter (Utiyama and Isahara, 2001). However, the results in Shi et al. (2007) suggest lexical cohesion based features may fare better for paragraphs when used in conjunction with other predictors.

Bag-of-words based similarity measures do not capture the fact that similar concepts can be expressed with different words, and so cannot give us the full story on segment cohesion. As such, similarity measures that account for this, such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) based topic modelling, have been applied to improve TextTiling and C99 style approaches (Choi et al., 2001; Riedl and Biemann, 2012). Thus, we would like to know if these representations can better capture lexical shifts associated with paragraph breaks. Similarly, we would like to get an idea of how well unsupervised Bayesian topic segmentation approaches work for paragraph segmentation. Like LDA based TextTiling, *BayesSeg* (Eisenstein and Barzilay, 2008) focuses on identifying distributional shifts rather than performing topic assignment, and thus makes a good candidate for our task. We would expect these approaches to work well if paragraphs are primarily characterized by changes in word distributions.

2.3. Prosodic Cues for Boundaries

Text-based studies indicate that lexical cues for paragraph boundaries can be quite domain specific. However, studies of prosody suggest that supra-sentential aspects of speech such as pitch, intensity and timing, may be a source of more robust boundary cues. A number of studies have shown that discourse boundaries are marked by similar prosodic features across speakers and domains. Such studies have consistently observed declination of pitch and intensity through the segment (Kreiman, 1982; Nakajima and Allen, 1993; Geluykens and Swerts, 1994; de Looze et al., 2015), reset to higher values at the beginning of new segment (Grosz and Hirschberg, 1992; Swerts, 1997; Tseng et al., 2006), and slower speaking rates near boundaries and pauses between boundaries (Lehiste, 1982; Smith, 2004; Zellers and Post, 2009).

The accumulated results suggest that prosodic boundaries share similar features across discourse levels. So we expect to see similar prosodic features at sentence internal phrase boundaries and at topic boundaries, albeit on a larger scale in the latter case. In line with this, Farrús et al. (2016) show that these declination, reset and timing properties generally hold over a large corpus of lectures with highly variable lexical/topical content. Boundary features, such as pauses and feature differences across sentences, have been shown to be useful for topic segmentation in the absence of lexical features (Shriberg et al., 2000; Levow, 2004a; Hirschberg and Nakatani, 1998). However, Farrús et al. (2016) also show that sentence intrinsic features such as sentence mean fundamental frequency (an acoustic correlate of perceived pitch) can also predict whether a sentence is paragraph final better than chance.

Boundary detectors based on non-lexical features are appealing in that they are less likely to be hurt by automatic transcription errors. In fact, Swerts and Geluykens (1993) show that human listeners can detect topical boundaries in instruction monologues where lexical content is obscured by a band-pass filter. Several studies have shown that models using only prosodic and acoustic features can be effective for broad topic segmentation (Hirschberg and Nakatani, 1998; Levow, 2004b; Tür et al., 2001; Hsueh, 2008; Wang et al., 2010; Zheng et al., 2012). However, the best performing segmentation approaches for spoken language generally use a combination of acoustic, prosodic and lexical features (Tür et al., 2001; Galley et al., 2003; Dielmann and Renals, 2007; Hsueh and Moore, 2007; Tsunoo et al., 2017). Thus, we expect that incorporation of prosodic features would be helpful for automatic paragraph segmentation.

2.4. Cue Words

The relationship between discourse related *cue words* or *cue phrases*, such as *'so'*, *'now'*, and *'okay'*, and discourse semantics has been extensively explored from a linguistic perspective (Schiffrin, 1987; Hirschberg and Litman, 1994; Knott and Dale, 1994, *interalia*). Moreover, discourse connectives, such as *'because'* and *'instead'*, can take arguments that span multiple sentences. Thus, we would expect the presence of such cue phrases to be useful for indicating multi-sentence constituents, which may in turn help define paragraph level segments. However, as mentioned above, previous studies have found features based on lists of discourse connectives to be less useful for predicting paragraph boundaries than cue words induced from the data during training (Filippova and Strube, 2006; Sporleder and Lapata, 2006). In fact, such induced

cues have been used to improve supervised and unsupervised lexical cohesion based topic segmenters (Galley et al., 2003; Rosenberg and Hirschberg, 2006; Hsueh et al., 2006; Eisenstein and Barzilay, 2008; Dowman et al., 2008).

Interestingly, examination of induced cue phrases suggests that generic discourse connectives such as *'so'*, *'and'*, and *'but'*, are more important in spoken language than written text (Eisenstein and Barzilay, 2008). Studies on the discourse uses of cue phrases have also shown interactions between cue phrases and prosody in terms of discourse interpretation. For example, Hirschberg and Litman (1994) show that the discourse structuring and sentential uses of words such as *'now'* correlate with different prosodic patterns. Thus, we expect the presence of lexical discourse cue phrases to be more predictive in speech segmentation, particularly in conjunction with prosodic features.

2.5. Feature Fusion

In general, we expect work on paragraph segmentation to shed light on how structure is cued within topical units via different feature modalities. Previous work on topic and sentence segmentation suggests that different types of lexical and acoustic characteristics will be relevant for identifying different levels of structure. However, how lexical and prosodic features should be combined remains an open question.

A number of previous studies use the concatenation of features across modalities as input to classifiers for topic segmentation, i.e., *early* or *feature level* fusion (Galley et al., 2003; Hsueh and Moore, 2007; Rosenberg and Hirschberg, 2006). However, models that integrate different feature families at later stages have improved segmentation performance on different datasets. This is often done by including boundary probabilities estimated from one model as input to another (*decision level fusion*). For example, Tür et al. (2001) find that decision level fusion of lexical and prosodic information within an HMM based story segmenter performed better than adding decision information from the lexical model to a prosody based decision tree boundary classifier. Similarly Dielmann and Renals (2007) find that allowing modalities to be processed independently and with different sampling frequencies, via a Dynamic Bayesian Network, improves meeting segmentation over a feature fused HMM.

Decision level fusion assumes that lexical and prosodic features present independent knowledge sources. This idea is counterintuitive given previous studies of how prosody can affect discourse meaning (Hirschberg and Litman, 1987, 1994). However, we should note that Tür et al. (2001), by design, only include prosodic cues that are deemed to be relatively unaffected by word identity. Moreover, since lexical knowledge is represented by boundary probabilities in their multimodal decision tree, direct interactions between lexical content and prosodic cues are not modelled. To take advantage of various subtle lexical and prosodic cues, we may need to combine evidence from different feature types at a relatively low level.

Given the previous discussion of suprasentential prosodic patterns and cohesion, we expect that modelling the temporal sequence of features is important for detecting boundaries. Recent work on multimodal sequence modelling tasks suggests that Recurrent Neural Networks (RNNs) are a good candidate for this job. They are also well suited for modelling different types of feature fusion. In terms of speech segmentation, Tilk and Alumäe (2016) show that adding pause features to hidden unit outputs from a text only RNN punctuation model improves performance for English and Estonian lecture data. Klejch et al. (2017) show that their lexical and acoustic RNN encoder-decoder model outperforms a lexical model for English punctuation. They also find that stochastically masking lexical inputs with acoustic features produces better results than simple feature concatenation. Similarly, Tsunoo et al. (2017) show that a hierarchical RNN model combining word embedding features and sentential acoustic features improves on the state-of-the-art methods in news story segmentation. Thus, structured fusion appears to work better than simply concatenating all features at the input level.

Overall, we expect RNN architectures to be expressive enough to allow us to evaluate different feature fusion strategies as well as the implications of sequence modelling for paragraphs. In the experiments that follow, we investigate the usefulness of RNN-based classifiers compared to AdaBoost, SVMs, and Multi-Layer Perceptrons (i.e. feed forward neural networks).

3. Experimental Setup

In the rest of this paper, we present automatic paragraph segmentation experiments on a large corpus of TED talks. The experiments that follow were designed to probe the following questions, as raised by the discussion above.

- What features are useful for paragraph segmentation?
- How can we make use of contextual information?
- How can we best use lexical and prosodic information?
- How well do our methods work on automatically transcribed speech?

We frame our task as a supervised learning problem. More specifically, we aim to predict whether a sentence precedes a paragraph boundary.[3] The experimental setup is described in the following sections.

3.1. Data

In this study, we build paragraph boundary detectors based on a set of 1492 TED (Technology, Entertainment, Design) talks published before 2014. These talks vary greatly in style and content. However, they are also generally known for being engaging and coherent. Thus, we expect them to be well structured and less likely to be affected by features of spontaneous speech like disfluencies or incomplete utterances. There is also less need for speakers to employ prosodic features for turn-taking purposes which may obscure the discourse structuring nature of these non-lexical features (Kreiman, 1982; Geluykens and Swerts, 1994; Gravano, 2009). The data set includes 1156 speakers who have a variety of English accents. As the talks span a wide variety of topics and genres, we also expect the lexical content of these talks to be less stylized than news broadcasts or agenda driven meetings. Thus, we would expect boundary indicating cues learned from this data to be more general.

Talks are 15 minutes long on average, so transcriptions generally benefit from paragraph segmentation. Each talk has been manually transcribed and includes punctuation and paragraph breaks. Transcribers are not given strict rules about paragraph structure. However, they do listen to the audio stream to help determine when paragraph breaks should occur.[4] We excluded talks consisting of only one paragraph (e.g. talks that are primarily musical performances or demonstrations). Altogether, the data set includes 210403 sentences and 33481 paragraphs with an average of 6 sentences per paragraph. We perform a random 90/10/10 split over the talks in the data set to form training, development and test partitions.

Sentence boundaries are detected based on transcribed punctuation using the Stanford CoreNLP sentence splitter (Manning et al., 2014). We use the same toolkit to obtain Part-of-Speech (POS) tags, parse trees, and co-reference information. We obtain word timings through Viterbi forced alignment using an automatic speech recognition system. Word timings are used to assign sentence boundary times. Given the aligned transcript, we extract various lexical and prosodic features as described in the following section and summarized in Tables 1, 2, and 3.

[3] These experiments extend the work presented in Lai et al. (2016).

[4] p.c. TED translation team.

3.2. Prosodic Features

We use Praat (Boersma, 2001) to extract Fundamental Frequency (F0) and intensity contours from the audio at 10 ms intervals. Here, F0 provides a measurable acoustic correlate of pitch, while intensity is a correlate of perceived loudness. To minimize errors in F0 estimation, we employed the method described in Evanini and Lai (2010) to set F0 parameter settings, as well as octave jump removal and linear interpolation through unvoiced segments. We normalize F0 and intensity values so that zero values represent speaker mean values for a talk. We subtract the speaker mean for intensity normalization, while F0 values are converted to log-scaled (semitone) values relative to speaker mean F0 value (Hz).

Based on our previous analysis of paragraph prosody (Farrús et al., 2016), we calculated aggregate statistics over the F0 and intensity contours of each sentence and the first and last words of each sentence (Table 1, STATS) including the mean, standard deviation, maximum, minimum, median, slope, quantiles, and range (difference between the 99th and 1st quantile values). Slope values are calculate using linear regression on the F0 and intensity values for the frame-level time sequence corresponding to the segment of interest. We also include the first five Legendre polynomial decomposition coefficients estimated over a given segment. The coefficients characterise the contour shape (e.g., bias, slope, convexity) and have been shown to be useful for detecting prosodic prominence (Kochanski et al., 2005). For timing features, we include the duration of the sentence, the number of words, the speaking rate (words per second), and the durations of pauses before and after the target sentence.

To better understand how different aspects of prosody indicate structure we experiment with different feature sets: (i) timing based features only (TIMING), (ii) sentence intrinsic F_0 and intensity features from the target sentence (TARGET), (iii) feature differences between the target sentence and the immediately preceding and following sentences, together with timing features (DIFF), (iv) timing, target sentence intrinsic features, and difference features, as well as TARGET features for sentences/words immediately preceding and following the target (PROSODY). To specifically understand the contribution of TIMING features, we also look at the full prosodic feature set without timing features. Based on (Farrús et al., 2016), we expect difference features to be the most indicative of paragraph boundaries. However, we would like to see if different model architectures are capable of making use of other prosodic features for this task. The feature sets are summarized in Table 1.

	Prosodic Features
STATS	*Calculated over F_0 and intensity contours:* • mean, std dev, max, min, range, slope • Quantiles(%): 1, 2.5, 25, 50, 75, 97.5, 99 • 1st-5th Legendre coefficients
TARGET	*For target sentence:* • STATS for whole target sentence • STATS for first and last words of target • target first and last word differences
TIMING	*For target sentence:* • Number of words, speaking rate • sentence duration • previous and next pause durations
DIFF	*Differences in* STATS *between:* • previous and target sentence • target and next sentence • target first word and preceding word • target last word and following word TIMING features
PROSODY	TARGET, TIMING, DIFFERENCE, previous/next sentence TARGET features

Table 1: *Summary of prosodic features and feature subsets.*

	Lexical Baseline Features
CW	1st, 2nd, 3rd word indicators, Cue phrase indicators (Knott, 1996)
BOW	Bag-of-Words indicators
LSX	average word entropy, sentence probability number of phrases, parse tree top level children, branching factor, parse tree depth, part-of-speech tag counts, number of words, relative position in doc, final punctuation, quote in previous, quote in target, incomplete quote,
BASELINE	CW, BOW, LSX

Table 2: *Summary of lexical baseline features based on Sporleder and Lapata (2006).*

3.3. Lexical Baseline and Cue Word Features

For the supervised lexical baseline (BASELINE), we extract features based on those used for paragraph segmentation of texts in Sporleder and Lapata (2006). The features fall into three categories: surface form, syntactic form, and language model based complexity features. Language models were estimated on the training partition using KenLM (1 to 5-grams) (Heafield, 2011) using modified Kneser-Ney smoothing without pruning. As in Sporleder and Lapata (2006), the language models were used to estimate average word entropy and sentence probabilities. The individual features are listed in Table 2, but we refer the reader to Sporleder and Lapata (2006) for more details. Lai et al. (2016) found that using the full baseline feature set performed better than models based on subsets of features, e.g. syntactic, surface, and language model based features. Since the current work is primarily interested in exploring the usefulness of discourse related features, here we focus on the full BASELINE set, cue words (CW), and bag-of-words (BOW) features.

From the lexical baseline features we identify two types of cue word related features (CW). We record the first three words of the sentence (i.e., three 1-hot encodings, excluding words that occur less than 100 times). We also include binary indicators for the presence of any cue phrases at beginning, middle and end of the sentence from the list in Knott (1996). We are specifically interested in the performance of these cue word features relative to bag-of-words features over the entire sentence (BOW, *k*-hot encoding). To look at how word identity features perform compared to other derived features in the baseline, we also report performance when combining cue word and bag-of-words features (CW+BOW), as well as the effect of their removal from the baseline set (LSX).

3.4. Lexical Cohesion Features

To examine the performance of lexical cohesion measures (COH), we look at the differences in topical and lexical similarity around potential boundary points based on various sentence vector representations (SENT VEC). These include Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Latent Semantic Analysis (LSA) (Deerwester et al., 1990), and TF.IDF representations of the transcript. We also extract sentence representations based on neural-network language models using the document vector approach of Le and Mikolov (2014) (D2V). The LSA and D2V models were trained using Gensim (Řehůřek and Sojka, 2010). LDA models were fit using Mallet (McCallum, 2002). In the training stage, individual talks were treated as documents. The words in each document were lemmatized and words that occurred in more than half of the talks were excluded. Numeric vector representations were assigned to individual sentences using these models (100 dimensional vectors for LDA, LSA, and D2V). Models were fit using only the training

	Lexical Cohesion Features
SENT VEC	LDA, LSA, D2V, TF.IDF (100 dim)
SIM	*Comparing windows before/after target:* • Smoothed similarity • TextTiling depth score *Target sentence similarity with:* • Previous 1-3 sentences • Next 1-3 sentences
CHAIN	*Lexical chain cohesion scores for:* • all lemmas, • non stopword lemmas, • lemmas with freq > 1, • co-referenced entities
COH	SIM for each SENT VEC type, CHAIN

Table 3: *Summary of lexical cohesion features. Vector similarity is measured using cosine distance. A window size of 3 sentences is used except where noted.*

partition of the data.

As in TextTiling (Hearst, 1997), we obtain similarity scores by summing sentence vectors falling inside the fixed windows (i.e., 3 sentences) before and after the target sentence, and then calculating the cosine similarity between these two vectors. We record the (moving-average) smoothed similarities across the boundary, as well as TextTiling depth scores. The latter measures the relative difference between the current similarity score and the closest 'peaks' in similarity to the left and right of the target sentence. A window size of 3 was used based on initial experiments using TextTiling for paragraph segmentation. We also include the cosine similarity of each target sentence vector with respect to the three previous and following sentences for a more local measurement of lexical change.

Besides topic model based features, we also measure cohesion based on lexical chains, i.e., word repetitions across sentences (Hearst, 1994; Galley et al., 2003). We calculate lexical chain cohesion scores (CHAIN) as the cosine similarity between the lexical chains in sentences before and after the potential boundary (i.e., rather than using all the lexical items in those windows). As in Galley et al. (2003), chains are weighted by the term frequency and compactness (i.e., log document length/chain length). We include separate features for chains based on all lemmas, lemmas that occur more than once, non-stopword lemmas, and chains based on automatically detected co-reference relations.

3.5. Baseline Classifiers

3.5.1. Unsupervised Baselines

For reference, we provide results based on well established unsupervised segmentation methods: BayesSeg (Eisenstein and Barzilay, 2008) on raw text input, as well as TextTiling and C99 based on sentence-level bags of words. In both cases we allow the segmenter to automatically determine the number of boundaries. We also use TextTiling and C99 with LDA, LSA, D2V and TF.IDF inputs. We use a block size of 3 sentences for TextTiling, and a mask size of 3 for C99. To help interpret their performance, we also give results for random and majority class segmentations (Niekrasz and Moore, 2010).

3.5.2. Baseline Supervised Classifiers

To compare with the supervised approach of Sporleder and Lapata (2006), we build classifiers using *AdaBoost* with decision stump estimators (Zhu et al., 2009). In this learning method, predictions are made via a linear combination of predictions from a number of potentially weak estimators (i.e., decision stumps). The weight assigned to each estimator is determined by their predictiveness with respect to the training input. In each learning iteration, examples that are misclassified in the previous round are given more weight (cf. boosting). Thus, we expect this method to work well in tasks where features are individually weak predictors.

We compare this with linear kernel *Support Vector Machine (SVM)* classifiers (Fan et al., 2008). These were previously used in Farrús et al. (2016) to investigate the predictiveness of prosodic features for identifying paragraph boundaries. However, in that work, combining all prosodic features produced worse performance than the best individual predictor. Since AdaBoost is designed to combine weak predictors effectively, we expect AdaBoost classifiers to have better performance than SVM classifiers.

AdaBoost and SVM classifiers were built using Scikit-Learn (Pedregosa et al., 2011). We tune hyper-parameters on the development set. For the SVMs, this is the regularization parameter C. For AdaBoost, we leave the learning rate at the default value (1.0), but tune the number of estimators (powers of 2 between 16 and 512).

3.5.3. Neural Network Classifiers

To investigate the ability of neural network based classifiers to model different feature interactions, we also perform experiments using *Multi-Layer Perceptrons (MLPs)* and *Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs)*. Each layer of an MLP consists of a number of hidden units (neurons) which are densely connected

to a vector of input features. Each hidden unit acts as a function whose output is the weighted sum of the input features. Non-linear functions may also then be applied to those outputs. This allows us to model complex, potentially non-linear interactions between input features.

In Recurrent Neural Networks (Elman, 1990), sequence information is modelled by feeding in the (weighted) hidden unit outputs generated from the input at the previous time step, alongside that of the current time step. This, in effect, allows the layer to retain memory of previous items in the sequence. This approach can also be used to model the sequence backward in time. Combining forward and backward RNNs over a sequence allows us to build up contextual representations in both time directions. This is potentially important for this segmentation task, since previous work suggests features of interest can straddle both sides of boundaries.

Several RNN variants have been proposed that add more structure to hidden units. Most prominently, in Long Short-Term Memory RNN (LSTM) hidden units are structured as memory cells with input, output, and forget gates (Hochreiter and Schmidhuber, 1997). This allows control over whether past context is used in computing the current outputs. Thus, we compare forward only LSTMs (LSTMs) with Bi-directional LSTMs (BLSTMs) in the following experiments to investigate the benefits of a more direct sequence modelling approach to incorporating contextual information. We refer the reader to Goldberg (2017) for a detailed introduction to MLPs and RNNs.

Neural network models were implemented using Keras (Chollet, 2015) and Tensorflow (Abadi et al., 2016). In these experiments, we use tanh activations and sigmoid output layers for both MLPs and BLSTMs. We use a grid search to select the number of hidden units (16 to 512 units) and layers (1 to 4) using the development set. We keep the number of hidden units the same for every layer. Since our outputs are binary, we train our networks with respect to cross-entropy loss using the RMSProp optimizer. To help prevent overfitting we employ dropout: 30% of the hidden unit outputs on each layer are randomly set to zero during training. We also perform early stopping based on development set loss.

3.6. Input Context Sequences

To get a better idea of how well context is used in different approaches, we perform experiments varying the amount of input context. More specifically, we add features from (1 to 4) previous and next sentences around every sentence in our corpus (zero padding at talk boundaries). This then forms a sequence of (3 to 9) input feature vectors which are used as input to the LSTM-based models. Model weights are trained using backpropagation over the network unrolled over the given sequence. The LSTM models make paragraph boundary predictions for every element in the input sequence. We use all such predictions when calculating losses during training, but we only use predictions for the target sentence when evaluating performance on the test set. For the MLP, AdaBoost and SVM models, we simply concatenate all vectors in the sequence to form one 'flat' input vector. The output then corresponds to the prediction for the center sentence in the input window. We treat the context size as a hyperparameter and tune this on the development set except where noted.

3.7. Feature Fusion

We investigate lexical and prosodic feature fusion at different levels by examining different BLSTM architectures as shown in Figure 1. For feature level fusion, we concatenate all features as input to a single BLSTM. For *decision level fusion*, we train separate lexical and prosodic BLSTMs, and make the final decision based on their separate class probability estimates. We also investigate whether learning separate feature representations for individual feature subsets before fusion improves performance. In this case, we combine the final hidden layer outputs, as opposed to output probabilities, from separate lexical and prosodic BLSTM models before making the final decision (*intermediate fusion*, cf. score fusion in Nandakumar et al. (2008)).

We use the best models for different feature sets as identified in the BLSTM feature level fusion experiments above. We compare cases where we make the final decision immediately after fusion, or whether we allow multiple BLSTM layers to intervene.

3.8. ASR Transcripts

To investigate the effect of transcription errors on our lexical features, we also apply our paragraph segmenters (trained on manual transcripts) to outputs of several ASR systems. ASR outputs for TED talks come from various systems developed at the University of Edinburgh (Bell et al., 2017). The systems display a wide variety of word error rates (10–50%). In the experiments below, we use the same sentence segmentation as for the manual transcripts. We apply a machine translation (text) based punctuation restoration system to obtain capitalization (to aid coreference resolution) and sentence internal punctuation (for lexical baseline features). In this case, prosodic features were extracted based on the ASR word timings.

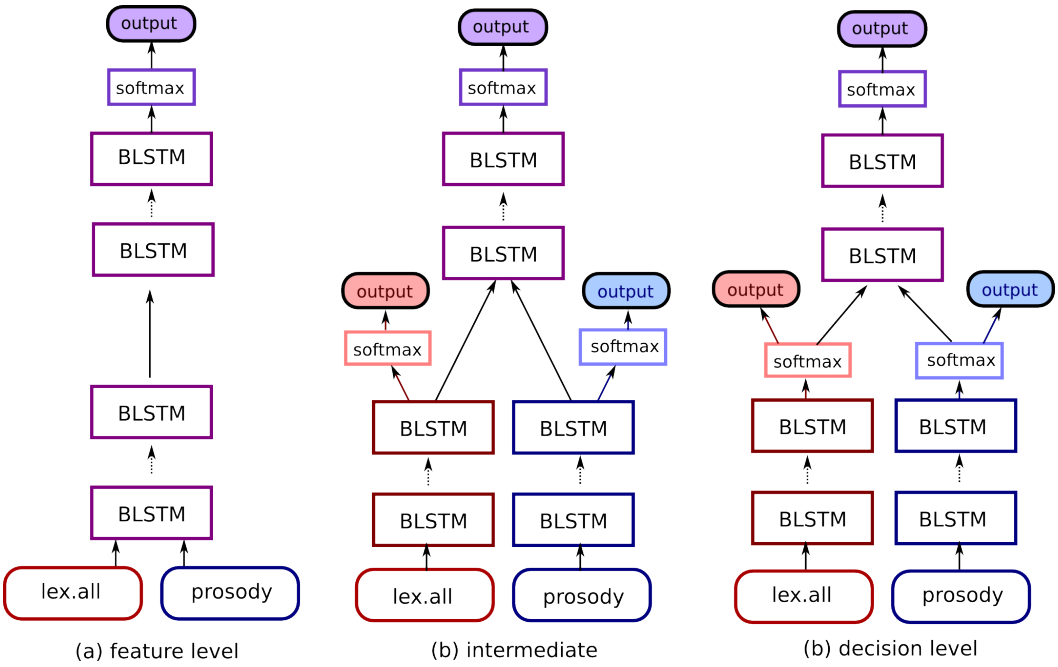

Figure 1: *Feature Fusion Architectures. Dotted arrows indicate that multiple layers of the same type, where the number of layers is treated as a hyperparameter and tuned on the development set.*

As the ASR systems used here were also trained on TED data, in these experiments we repartition the data to use the 27 talks in the ASR test sets (dev2010, tst2010, tst2011) as our paragraph segmentation test set. All other talks were split 90/10 into training and development data. Paragraph segmenters were retrained using the hyperparameter settings selected in the experiments with manual transcripts.

3.9. Evaluation Metrics

To evaluate our results we initially use standard segmentation metrics: P_k (Beeferman et al., 1999) and WindowDiff (WD) (Pevzner and Hearst, 2002). Both metrics measure segmentation error using a sliding window (size k) through a document. For P_k, a penalty of 1 is added if a boundary is predicted for a no-boundary window or vice versa. For WD, a penalty of 1 is added if the predicted number of boundaries does not match the ground truth. The summed penalties are normalized by the total number of windows to produce an error probability, with 0 indicating a perfect segmentation (i.e., lower P_k and WD scores are better).

These standard metrics are known to be biased towards segmentations with very few predicted boundaries or edge clumping. Thus, following Niekrasz and Moore (2010), we also report k-κ, a version of P_k which is explicitly corrected for chance agreement. We also extend the document sequence with k-1 zeros (no boundary) at either end to ameliorate the edge bias problem. We use k=3 following the standard practice of using half the average segment length for the dataset. For k-κ, scores of -1, 0, and 1 represent perfect disagreement, chance and perfect agreement respectively. To examine the significance of differences between models, we use Paired Permutation Tests over k-κ (Yeh, 2000; Smith, 2011), with $p < 0.05$ as the criteria for statistical significance. Any mention of significant differences in the following refers to a statistically significant difference according to this test.

4. Results

In the following, we report results of various approaches for automatically predicting paragraph boundaries. The evaluation metrics we show represent TED talk test set results only (cf. Section 3.1).

4.1. Text-Based Unsupervised Baselines

Table 4 shows various unsupervised baseline results for various vector representations of the text. As expected, k-κ is close to zero for both of these baselines. For both C99 and TextTiling, the TF.IDF based approaches perform better than LDA, LSA, D2V and Bag-of-Words (BOW) based approaches. C99 results are generally better than TextTiling in terms of k-κ, but worse in terms of WD. This reflects a higher recall but also a higher false positive rate for the former. Overall, the Bayesian topic modelling approach

Classifier	Features	P_k	WD	k-κ
Majority	n/a	0.41	**0.41**	0.00
Random	n/a	0.45	0.50	0.03
TextTiling	TF.IDF	0.40	0.41	0.11
	LDA	0.42	0.44	0.09
	LSA	0.43	0.45	0.09
	D2V	0.44	0.46	0.09
	BOW	0.46	0.48	0.04
C99	TF.IDF	**0.39**	0.41	0.12
	D2V	0.43	0.44	0.10
	LDA	0.45	0.49	0.10
	LSA	0.45	0.51	0.10
	BOW	0.47	0.57	0.06
BayesSeg	text	0.40	0.50	**0.16**

Table 4: *Unsupervised baseline test set results. The majority class baseline predicts no boundaries anywhere. For P_k and WD scores, lower scores indicate better performance. For k-κ, higher scores indicate better performance.*

(BayesSeg) performed the best in terms of k-κ, although it appears to suffer in WD terms from a higher false positive rate. Overall, we see that these unsupervised lexical cohesion based methods can detect some boundaries, and abstract representations help compared to bags-of-words. However, the observed performance is generally low, supporting the hypothesis that lexical similarity is a weak indicator of paragraph structure.

4.2. Feature Set Comparisons

In the following, we present results examining the performance of prosodic and lexical feature subsets, as well as combinations of lexical and prosodic features. Given the problems with P_k and WD outlined above, we focus on k-κ results in our comparisons. Differences in k-κ in the results discussed in the text below were found to be statistically significant ($p < 0.05$), except where noted.

4.2.1. Prosodic Features

Results from different prosodic feature sets are shown in Table 5. Using the full prosodic feature set (PROSODY) performed the best across all classifier types. Prosodic difference features (DIFF) are clearly an important part of this. However, target sentence intrinsic features (i.e., TARGET) are also useful for this task in conjunction with difference features (i.e., DIFF < PROS ALL). Removing TIMING features from the full prosodic feature set reduces performance considerably, but again there is clearly still some useful information in the non-timing features. This confirms the idea that multiple aspects of sentence prosody are required for this task.

Input	ADABOOST	BLSTM	LSTM	MLP	SVM
TARGET	0.12	0.16	0.06	0.12	0.01
TIMING	0.15	0.20	0.15	0.19	0.03
DIFF	0.19	0.24	0.21	0.21	0.11
PROSODY	**0.22**	***0.27***	**0.22**	**0.25**	**0.12**
w/o TIMING	0.10	0.15	0.14	0.12	0.07

Table 5: *Prosodic features: Test set results for the full prosodic feature set* PROSODY *and subsets:* TARGET *sentence F_0 and intensity features,* TIMING *features including pausing, and* DIFF*erences between the target and previous/next sentences. The last row shows the effect of removing the* TIMING *features from the full prosodic feature set. Larger k-κ values indicate better performance.*

Using the BLSTM with the full prosodic feature set (PROSODY) performs the best overall. As expected, the neural network approaches generally perform better than the baseline AdaBoost and SVM classifiers. It appears that the prosodic difference features were better utilized by the SVM than duration (i.e., TIMING) based features or sentence intrinsic features, even when context was available. These results suggest that some notion of feature composition is necessary to get the most out of non-difference features for this task. The forward LSTM results are always worse than the bidirectional version and the context dependent MLP. The difference is especially pronounced for the model that doesn't include any contextual DIFFERENCE features (i.e., TARGET). The fact that the MLP is able to recover relevant information from context where the forward LSTM cannot, indicates that some post-target information is necessary in this task setup.

4.2.2. Lexical features

Table 6 shows the results for experiments with lexical feature subsets. The lexical baseline models perform better than prosody based models for all classifiers except the LSTM. In general, lexical cohesion features perform better than chance, but are still quite weak compared to the other feature sets. Adding cohesion features (COH) to the baseline features generally improves performance except for the BLSTM, where the results are not statistically significantly different (LEX VS BASELINE). This suggests that the BLSTM is able to capture information similar to (or at least as informative as) the lexical cohesion features through sequence modelling.

Lexical cohesion features on their own do not perform as well as cue word (CW) and bag of words (BOW) features except in the forward LSTM case. As with the prosodic features, not being able to access features from across the potential boundary, particularly cue words, handicaps the predictiveness of this model. Removing cue word and bag

of words features from the lexical baseline set (i.e., LSX) degrades performance significantly. This indicates that these lexical identity features are important components of the lexical baseline feature set. However, the additional surface, syntax and language model features (included in LSX) also add something beyond the identity features alone (BASELINE > CW+BOW).

Overall, the fact that lexical cohesion features do not perform as well as cue word features supports the idea that discourse oriented cues are more important than lexical cohesion for this level of segmentation. We also see that some models are able to integrate cue word and bag-of-words features better than others. As for the prosody experiments, the BLSTM based models perform the best followed by the MLP, AdaBoost, SVM, and lastly the forward LSTM. This indicates that the type of transforms afforded by the neural networks are useful for this task, but access to the right type of context is required.

4.2.3. Lexical and Prosodic features

Table 7 shows the performance of classifiers trained on combined sets of lexical and prosodic classifiers. Using the full set of combined prosodic and lexical feature sets (LEX PROS) improves performance compared to individual lexical or prosodic models for each classifier type. We see a similar improvement when just combining cue word and bag-of-words features with prosodic information (CWB PROS), although performance here is worse than for the full feature set. Overall, the results support the idea that prosodic and lexical features contribute complementary information for this task.

The gain from adding prosodic features appears greatest for the LSTM model. This is likely due to the fact that the prosodic features include next sentence features, and thus provide useful additional lookahead type information. In terms of classifiers, we see a now familiar ordering between model types: BLSTM, MLP, AdaBoost, LSTM, SVM. In this case the difference between MLP and AdaBoost is not statistically significant, nor is the difference between the SVM and LSTM. This, once again, reinforces the idea that composition of features in time is important for this task.

4.3. The Effect of Context

As discussed in Section 3.6, we varied the amount of context available to classifiers to investigate how access to context affects performance. In the experiments above, we only reported results for the context size that produced the best results on the development set. To give a more complete picture, Figure 2 shows the test set results when varying amounts of context are provided as input (or training sequence length for (B)LSTMs). Here, we see that adding just features from the previous and following sentence to the target sentence results in a huge jump in performance for non-reccurent lexical models. However, the prosodic models don't gain as much from including additional context. Note that the full prosodic feature set includes information about the immediately preceding and following sentences. The lack of improvement from adding more context indicates that the most useful prosodic information for this task is quite close to the potential boundary.

The superior performance of the BLSTM compared to the MLP indicates that explicit mechanisms for sequence modelling employed in the recurrent network are important here. It appears that the MLP is less able to model lexical and prosodic interactions as the context size grows. We can also note that the forward LSTM results remain much more stable with increasing context size with lexical features generally behaving quite poorly. As in the results above, we see that both forward and backwards sequence modelling is important for this task.

4.4. BLSTM-based Fusion

The results described above all employ feature level fusion: i.e, the input is the concatenation of all features. For neural networks, the dense connections between layers basically allow interactions between input features. However, it is possible that allowing these sorts of interactions is unnecessary. Previous work has suggested that we can treat lexical and prosodic features as independent information streams (cf. Section 2.5). Training separate neural networks for feature subsets essentially removes connections between feature blocks, reducing the number of model parameters. Thus, if we can treat specific feature sets as independent, we may be able to fit the data more efficiently.

We investigate this by looking at models that perform two types of late fusion: decision level fusion and intermediate fusion (Figure 1). We also compare different partitions of the input features. In the first case, we use separate models for lexical features without word identity features, (LSX), the combined cue words and bag of words features (CW+BOW), lexical cohesion features (COH), and the full prosodic feature set (PROSODY). In the second case, we simply separate all lexical (LEX) and prosodic features (PROSODY). We use the best models for different feature sets as identified in the experiments above, so model components differ in number of hidden units and layers before fusion.

Table 8 shows the results for different feature fusion strategies. Here, our previous *feature* level fusion model

Input	ADABOOST	BLSTM	LSTM	MLP	SVM
COH	0.07	0.08	0.07	0.08	0.07
CW	0.14	0.22	0.07	0.17	0.14
BOW	0.14	0.23	0.09	0.14	0.14
LSX	0.15	0.23	0.09	0.18	0.13
CW+BOW	0.18	0.29	0.07	0.23	0.19
CW+BOW+LSX (=BASELINE)	0.23	***0.34***	0.12	0.27	0.24
CW+BOW+LSX+COH (=LEX)	**0.26**	***0.34***	**0.18**	**0.28**	**0.25**

Table 6: *Lexical features: Test set results (k-κ).* BASELINE *is the feature set used in Sporleder and Lapata (2006).* LEX *represents all* BASELINE *and* COH*erence features.*

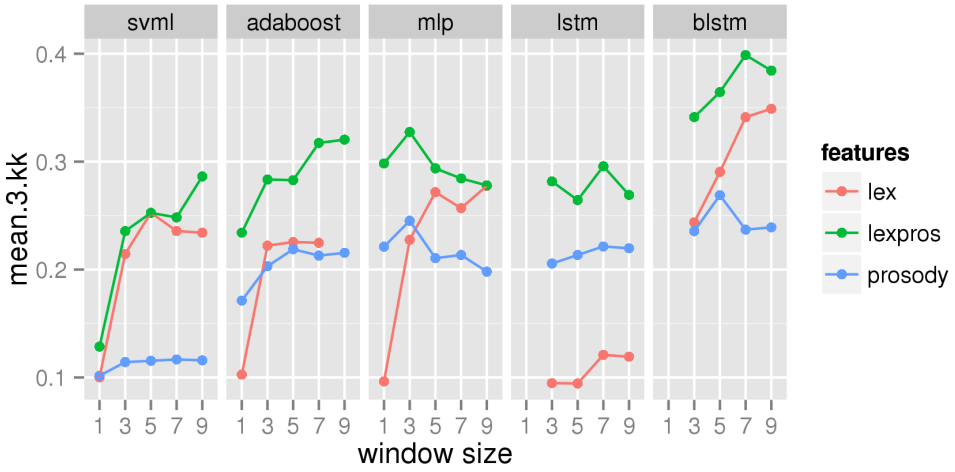

Figure 2: *Varying the available context: Test set mean k-κ for models trained using best dev set hyperparameters. For a window size of 3, for example, the input to the SVM, AdaBoost, and MLP consists of target sentence features, plus the same features for the immediately previous and following sentences*

Input	ADABOOST	BLSTM	LSTM	MLP	SVM
CWB PROS	0.29	0.36	0.25	0.29	0.25
LEX PROS	**0.32**	***0.38***	**0.30**	**0.33**	**0.29**

Table 7: *Combining Lexical and Prosodic features: Test set results (k-κ, feature level fusion).* CWB PROS: *cue words, bag of words and all prosodic features.* LEX PROS: *all lexical features and all prosodic features.*

(cf. Table 7) consistently outperforms *decision* level fusion, where we simply use a sigmoid decision layer on top of output probabilities. However, we get better results when we perform *intermediate* fusion, where the fusion layers take hidden unit outputs of the separate lexical and prosodic models as inputs to the final decision layer. That is, allowing features derived from the separate lexical and prosodic models to interact improves performance.

However, we can improve both decision level and intermediate fusion by adding additional BLSTM layers between the individual models and the overall (*decision+blstm*) decision level. This again highlights the sequential dependencies in prediction. We see a small but statistically significant gain using the intermediate fusion strategy (*intermediate+blstm*). This gives us our best results overall. Both of these models outperform the feature fusion model that matches their overall depth (i.e., 8 layer feature fusion). Thus, allowing composition of prosodic and lexical features before the decision layer is generally preferable. Furthermore, additional partitioning of the input feature space reduces performance. This indicates that some interactions between lexical features used here need to happen early in the model structure.

We also consider earlier fusion where we take separate 1 layer paragraph boundary models and fuse them with BLSTM layers (cf. Lai et al. (2016)). The results in Table 9 show that intermediate fusion performs better than decision and feature level fusion when we do not include more fusion layers. We obtain better performance if we allow more BLSTM layers after fusion. However, Table 9 also shows that input feature level fusion with a 5 layer model

Fusion Type	Fusion Layers	μ(k-κ)
LSX, CW+BOW, COH, PROSODY		
Decision	0	0.28
Intermediate	0	0.36
Decision+blstm	4	0.40
Intermediate+blstm	4	0.41
LEX, PROSODY		
Decision	0	0.34
Intermediate	0	0.38
Decision+blstm	4	0.41
Intermediate+blstm	4	**0.42**
Feature (Table 7)	4	0.38
Feature	8	0.39

Table 8: *Test set results for different fusion types and different partitions of the full input feature set (mean k-κ). We train separate BLSTM models for the different input feature types before fusion. The number of fusion layers indicate the number of BLSTM layers combining the different feature types not including the final decision layer.*

Fusion type	Layers	μ(k-κ)
Decision	1	0.22
Feature	1	0.26
Intermediate	1	0.29
Intermediate+blstm	1+4	0.34
Feature	5	**0.38**

Table 9: *Test set results for fusion after 1 layer lexical and prosodic BLSTMs (mean k-κ).*

outperforms one where the first layer separates lexical and prosodic features. This indicates that several layers are necessary to obtain a good representation for late fusion.

4.5. Application to ASR outputs

While the full lexico-prosodic model performs best on the manual transcripts, we'd like to see how well classifiers built on different feature sets perform on ASR outputs. Figure 3 shows how *k-κ* varies with Word Error Rate (WER) for various models (feature level fusion). As we would expect, the graph shows a steady decrease in performance for the lexical model with increasing WER. However, the addition of prosodic features ameliorates the degradation caused by ASR errors. This occurs even though the prosodic model performs relatively poorly on its own. ASR errors are likely to result in incorrect word timing features which affect aggregate prosodic features over words and sentences, as well as pause information. Thus, we observe a large decrease in the prosodic model compared to the manual transcription case. So, it appears that it is quite hard to decouple prosodic and lexical features if the former depend on word timings. Nevertheless, degraded prosodic features can still help out degraded lexical features.

Interestingly, the model using just cue words, bag of words, and prosodic features (CWB PROS) outperforms the superset lexical and prosodic (feature fusion) model (LEX PROS). The full lexical model includes features derived from other language processing tasks, such as syntactic parsing, which are likely to degrade with ASR errors. In contrast, CW and BOW features simply rely on word identity and so are less sensitive to this problem.[5] This may account for the CWB PROS outperforming the LEX PROS model. However, it still appears that the cue word/bag-of-words model (CW BOW) results degrade less quickly than the full LEX results. In both cases, the prosodic information improves performance compared to corresponding models that don't include prosodic information.

Results corresponding to different fusion methods are shown in Figure 4. This shows that decision level fusion performs better than intermediate fusion on the manual transcriptions than for the ASR test set. However, intermediate fusion performs better on the ASR output. This may be because the overall confidence of individual lexical and prosodic models is reduced when dealing with ASR outputs. Both intermediate fusion and decision level fusion perform better than feature level fusion. This, once again, suggests that performing some separate abstraction over lexical and prosodic features is beneficial. However, combining them in terms of abstract representations (hidden

[5] It is worth noting that performance of cue word/bag-of-word features on this small test set appear to be quite a lot better than for the separate test set used in the experiments above, although the feature set performance ordering is otherwise the same on manual transcripts.

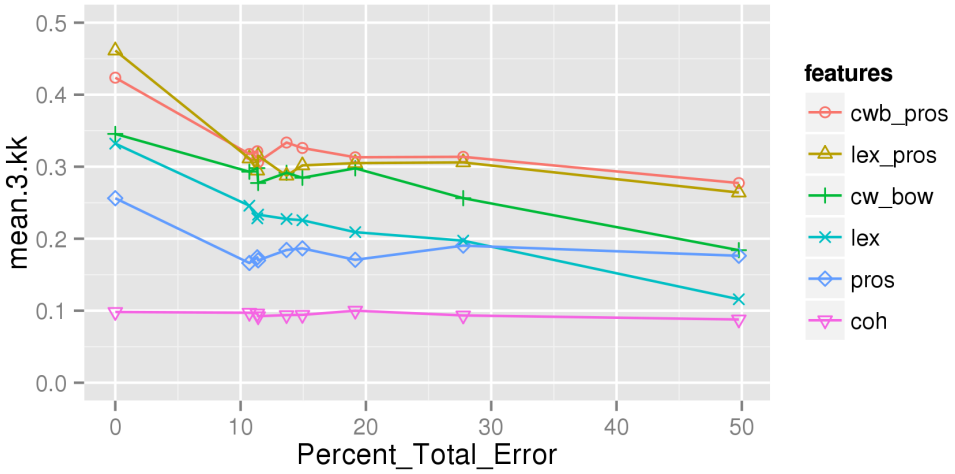

Figure 3: *Results for the ASR test set using ASR transcripts with feature level fusion models: Word error rate versus mean k-κ. Zero percent total error results correspond to predictions made using manual transcripts as input.*

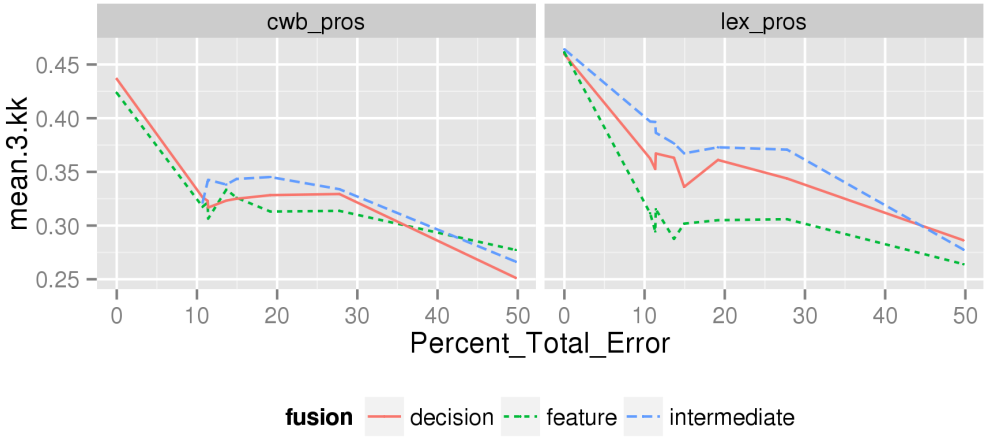

Figure 4: *Results for the ASR test set using ASR transcripts for different fusion types: Word error rate versus mean k-κ.*

layer outputs) is more robust than using decision outputs.

It is important to note that these experiments focus on the effect of ASR errors on this task given gold-standard sentence timings. This was done to in order to keep things consistent with the experiments performed on manual transcription (i.e., 0% WER): ASR errors effect punctuation restoration, which can in turn change sentence segments. This makes evaluation with respect to the gold standard less straightforward and would require us to use a modified evaluation metric. Since the focus of the current work is on establishing the usefulness of prosodic and lexical features for this task, we defer a detailed investigation into all upstream components for future work.

However, at this stage we can note that punctuation restoration is still an area of ongoing research, with previous work on TED talks report overall F1-scores 61.4%, and 71.4% when restricted to period restoration Tilk and Alumäe (2016). So, we would expect this to effect downstream paragraph segmentation negatively. Beyond this, work on spontaneous conversational speech has noted difficulties in translating the concept of punctuated sentence from text to spoken utterances (Strassel, 2003). Similarly, the TED talks don't often exhibit the types of disfluencies that are the hallmark of conversational speech (Shriberg, 2001). As we saw previously, timing features like pausing are predictive of paragraph boundaries. However, the presence of pauses related to disfluencies make the interpretation of pauses more complicated (and similarly so for sentence segmentation). So, we would expect to see more reliance on lexical cues in this case, although more long range prosodic features may still be able to compensate for this. We would not expect our models trained on fluent speech to be able to generalize to disfluent speech without some adaptation. A model that considers word-level sentence segmentation probabilities obtained, for example, in a multitask training scenario, may be more appropriate for discourse segmentation of spontaneous speech. In general, further work is required to better understand the relationship between spoken utterances and larger discourse segments represented by paragraphs, especially in presence of disfluencies.

5. Discussion

The experiments described above indicate that cue words and prosody are better indicators of paragraph structure than measures of lexical cohesion. This supports the idea that discourse oriented surface cues, e.g. word identities, are more important for this level of segmentation than changes in word distributions. Although lexical cohesion features did improve performance relative to the lexical baseline for most classifiers we experimented with, they appeared redundant in the corresponding BLSTM experiments. So, while lexical/topic similarity measures do reflect some aspects of paragraph structure, they are far from sufficient for capturing paragraph cohesion.

In line with Sporleder and Lapata (2006), our results indicate that combining a large number of individually weak lexical and prosodic predictors is necessary for this task. Moreover, we can improve performance by allowing interactions between lexical and prosodic signals. The classification methods we investigated vary to the extent that they allow such interactions. Machine learning methods such as AdaBoost can do this to some extent by learning a weighted combination of estimators. However, neural network models can go further by using the hidden unit structure to compose feature values through a series of linear and non-linear functions. The benefits of including prosodic features are particularly evident as the error rate increases in automatic transcripts.

In fact, the experiments show that lexico-prosodic models benefit from late fusion, where separate lexical and prosodic models, in effect, act as feature extractors. Fusing the hidden layer outputs tends to work better than decision level fusion, particularly for ASR transcripts. This indicates that some extra abstraction over the current input feature set is required. A benefit of delaying modality fusion in this way is that it allows models of different modalities to be trained separately. For example, the lexical component could take advantage of text without corresponding speech data. However, given domain differences observed in previous text studies, further work is necessary to see how well text oriented paragraph cues transfer to speech.

Making more contextual information available to SVM, AdaBoost, and MLP based classifiers improved performance. However, the BLSTM models outperform these methods, showing the benefits of structured composition of features across time. Experiments varying input sequence length support previous findings that the most useful prosodic features for segmentation occur very close to potential boundaries (Shriberg et al., 2000). However, the overall results also support the idea that lexico-prosodic interactions indicate discourse structure. In the current work, we only extract the prosodic word level features for the first and last word of a sentence, so there is only a direct correspondence with the first cue word feature. However, incorporating word or subword prosodic features together with word representations, such as word embeddings, for all words in the sentence may better model discourse structure related interactions.

In this vein, combining (sub-)word features may also help us to obtain more useful and compact sentence representations using, for example, autoencoder based methods (Li et al., 2015). Similarly, we plan to investigate the usefulness of lexical and prosody based attention mechanisms for this task (Luong et al., 2015; Bahdanau et al., 2014). Learning paragraph segmentation as a joint task with sentence segmentation (Tilk and Alumäe, 2016; Klejch et al., 2017) or discourse parsing (Ji and Eisenstein, 2014; Braud et al., 2016) may also be helpful for learning appropriate lexical and prosodic representations. This may, in turn, allow us to better model supra-sentential sequences. However, even without these potential enhancements, our current results clearly show the benefits of incorporating lexical and prosodic features before the decision level and of sequence modelling.

Bayesian generative models provide an alternative approach to integrating knowledge sources and adapting to different styles for unsupervised segmentation (Dowman et al., 2008; Nguyen et al., 2012; Du et al., 2013). In the current work, we did not explore any of these methods besides applying the BayesSeg baseline. It would be interesting to see whether integrating prosodic features into these sorts of models at the word level can help capture the more subtle shifts represented by paragraphs. In fact, it appears that little work has been done on incorporating non-lexical speech features into topic models. The model presented in Dowman et al. (2008) shows how to incorporate discourse features such as speaker activity and overlap into a probabilistic model of meeting segmentation. However, this approach assumes that features are independent of each other and feature distributions depend only on the segment boundary variable. In contrast, our current work indicates that paragraph segmentation requires incorporation of sequence information and modelling of interactions between features. Neural variational inference techniques for learning latent variable models, like topics (Miao et al., 2017), may provide a promising avenue to further explore Bayesian generative models for multimodal paragraph segmentation using recurrent neural networks. Similarly, the paragraph segmentation task may also benefit from incorporating aspects of recent work on topic segmentation in task-oriented dialogue using reinforcement learning (Takanobu et al., 2018). We leave investigation of this for future work.

Several studies have found that, while human agreement for discourse segmentations is far from perfect, some boundaries are clearly more popular than others (Stark, 1988; Ji, 2008; Passonneau and Litman, 1993a). In order to understand what types of boundaries our neural network models are finding, we need a better understanding of the types of thematic units they present. For example, discontinuities may be better understood in terms of re-orientations of participants, perspective, location, and time. Since documents need to maintain multi-paragraph coherence, these discontinuities may not be very sharp (Giora, 1986).

At this time, human performance on our paragraph segmentation task is unknown. Most recent works on topic and paragraph segmentation have focused on prediction of automated methods with respect to pre-existing (author) text segmentations as gold standards, rather that human performance. The closest study is Niekrasz (2012), who calculates k-κ of 0.58 for speaker intention based segmentation of Pear story narratives (mean over 7 annotators). Our best performance is still some ways off from that level. However, the results are not directly comparable, e.g., Niekrasz used a window size of 1 for comparability to Cohen's κ agreement measure. Thus, further investigation of human agreement on paragraph segmentation with respect to different types of shifts on this data set will be necessary in future work.

6. Conclusion

This paper investigated automatic paragraph segmentation using lexical and prosodic features on a highly diverse set of TED lectures. Our experiments show that cue word and bag-of-words features performed better than lexical cohesion measures drawn from work on automatic topic segmentation. This indicates that lexical distribution shifts only tell part of the story of what makes the lexical content in a paragraph cohesive. The performance of cue word/bag-of-word features suggests that specific words are indicative of this level of discourse structure. Similarly, the structure indicating aspects of prosody can help make up for errors in ASR transcripts. In fact, in experiments on ASR outputs, classifiers based on cue word, bag-of-words, and prosodic features outperformed the full feature set as word error rate increased.

Overall, these experiments highlight the need to combine many weak predictors to make good boundary decisions. Moreover, how features are combined can have a large impact on results. Approaches that can model interactions between low level features worked better. Overall, the best performance came from BLSTM models using the full lexical and prosodic feature sets. We observe that the inclusion of sequential information both forwards and backwards in time resulted in significant gains for this task. Late fusion of lexical and prosodic features provided the best results overall. Intermediate fusion of hidden layer

outputs worked better than decision level fusion, especially for automatically transcribed talks.

Future work will focus on mapping between representations learned by neural network models to the different types of thematic discontinuity represented by paragraph breaks. We will also investigate richer neural network based sentence representations that incorporate word and sub-word lexical and prosodic features. Further work is also required to determine how well this approach generalizes to other spoken language genres (e.g. news broadcasts), and, similarly, representations learned from written text transfer to this task.

Acknowledgments

The second author was funded from the *EU's Horizon 2020 Research and Innovation Programme* under the GA H2020-RIA-645012 and the Spanish Ministry of Economy and Competitivity *Juan de la Cierva* program. The other authors were funded by the University of Edinburgh.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic Detection and Tracking Pilot Study Final Report. *Proceedings of the Broadcast News Transcription and Understanding Workshop (Sponsored by DARPA)*, February 1998. URL http://repository.cmu.edu/compsci/341.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Doug Beeferman, Adam Berger, and John Lafferty. Statistical Models for Text Segmentation. *Machine Learning*, 34(1-3):177–210, February 1999. ISSN 0885-6125, 1573-0565. doi: 10.1023/A:1007506220214.

P. Bell, P. Swietojanski, and S. Renals. Multitask Learning of Context-Dependent Targets in Deep Neural Network Acoustic Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(2):238–247, February 2017. ISSN 2329-9290. doi: 10.1109/TASLP.2016.2630305.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003. ISSN ISSN 1533-7928. URL http://www.jmlr.org/papers/v3/blei03a.html.

Paul Boersma. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345, 2001.

Chloé Braud, Barbara Plank, and Anders Søgaard. Multi-view and multi-task training of RST discourse parsers. In *26th International Conference on Computational Linguistics (coling)*. Association for Computational Linguistics, 2016. URL https://chloebt.github.io/publications/braudPlankSoegaard_coling16.pdf.

Freddy Y. Y. Choi. Advances in Domain Independent Linear Text Segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 26–33, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=974305.974309.

Freddy Y.Y. Choi, Peter Wiemer-Hastings, and Johanna D. Moore. Latent Semantic Analysis for Text Segmentation. In *Proceedings of Empirical Methods in Natural Language Processing*, 2001. URL http://www.cs.cornell.edu/home/llee/emnlp/papers/choi.pdf.

François Chollet. Keras. https://github.com/fchollet/keras, 2015.

Céline de Looze, Irena Yanushevskaya, Andy Murphy, Eoghan O'Connor, and Christer Gobl. Pitch Declination and Reset as a Function of Utterance Duration in Conversational Speech Data. In *Proceedings of Interspeech 2015*, 2015.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science; New York, N.Y.*, 41(6):391–407, September 1990. ISSN 0002-8231. URL http://search.proquest.com/docview/1301252034/citation/2EF3417906F414APQ/1.

A. Dielmann and S. Renals. Automatic Meeting Segmentation Using Dynamic Bayesian Networks. *IEEE Transactions on Multimedia*, 9(1):25–36, January 2007. ISSN 1520-9210. doi: 10.1109/TMM.2006.886337.

M. Dowman, V. Savova, T. L. Griffiths, K. P. Kording, J. B. Tenenbaum, and M. Purver. A Probabilistic Model of Meetings That Combines Words and Discourse Features. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1238–1248, September 2008. ISSN 1558-7916. doi: 10.1109/TASL.2008.925867.

Lan Du, Wray L. Buntine, and Mark Johnson. Topic Segmentation with a Structured Topic Model. In *HLT-NAACL*, pages 190–200, 2013. URL http://www.aclweb.org/website/old_anthology/N/N13/N13-1019.pdf.

Jacob Eisenstein and Regina Barzilay. Bayesian Unsupervised Topic Segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 334–343, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1613715.1613760.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Keelan Evanini and Catherine Lai. The importance of optimal parameter setting for pitch extraction. *The Journal of the Acoustical Society of America*, 128(4):2291–2291, October 2010. ISSN 0001-4966. doi: 10.1121/1.3508047.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

Mireia Farrús, Catherine Lai, and Johanna D. Moore. Paragraph-based prosodic cues for speech synthesis applications. In *Proceedings of Speech Prosody 2016*, pages 1143–1147, Boston, MA, USA, 2016.

Katja Filippova and Michael Strube. Using Linguistically Motivated Features for Paragraph Boundary Identification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 267–274, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 978-1-932432-73-2. URL http://dl.acm.org/citation.cfm?id=1610075.1610114.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan

Jing. Discourse Segmentation of Multi-party Conversation. In *Proc. ACL*, pages 562–569, Stroudsburg, PA, USA, 2003.

Ronald Geluykens and Marc Swerts. Prosodic cues to discourse boundaries in experimental dialouges. *Speech Communication*, 15(1–2):69–77, October 1994. ISSN 0167-6393. doi: 10.1016/0167-6393(94)90042-6. URL http://www.sciencedirect.com/science/article/pii/0167639394900426.

Maria Georgescul, Alexander Clark, and Susan Armstrong. A Comparative Study of Mixture Models for Automatic Topic Segmentation of Multiparty Dialogues. In *IJCNLP*, pages 925–930, 2008. URL https://www.aclweb.org/anthology/I/I08/I08-2133.pdf.

Rachel Giora. Principles of segmentation in the literary text. The case of the formally unsegmented text. *Journal of Literary Studies*, 2(2):15–28, July 1986. ISSN 0256-4718, 1753-5387. doi: 10.1080/02564718608529790. URL http://www.tandfonline.com/doi/abs/10.1080/02564718608529790.

Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

Susan R. Goldman, Elizabeth U. Saul, and Nathalie Coté. Paragraphing, reader, and task effects on discourse comprehension. *Discourse Processes*, 20(3):273–305, November 1995. ISSN 0163-853X. doi: 10.1080/01638539509544942. URL http://dx.doi.org/10.1080/01638539509544942.

Agustín Gravano. *Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue*. PhD thesis, Columbia University, 2009.

B.J. Grosz and Julia Hirschberg. Some intonational characteristics of discourse structure. In *Proceedings of ICSLP*, pages 429–432, 1992.

Kenneth Heafield. KenLM: faster and smaller language model queries. In *Proc. EMNLP Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July 2011.

Marti A. Hearst. Multi-paragraph Segmentation of Expository Text. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 9–16, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. doi: 10.3115/981732.981734. URL http://dx.doi.org/10.3115/981732.981734.

Marti A. Hearst. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64, March 1997. ISSN 0891-2017.

J. Hirschberg and D. Litman. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3):501–530, 1994.

Julia Hirschberg and Diane Litman. Now Let's Talk About Now: Identifying Cue Phrases Intonationally. In *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*, ACL '87, pages 163–171, Stroudsburg, PA, USA, 1987. Association for Computational Linguistics. doi: 10.3115/981175.981198. URL http://dx.doi.org/10.3115/981175.981198.

Julia Hirschberg and Christine H. Nakatani. Acoustic indicators of topic segmentation. In *Proceedings of ICSLP 1998*, Sydney, 1998.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Pei-Yun Hsueh. Audio-based unsupervised segmentation of multiparty dialogue. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5049–5052, March 2008. doi: 10.1109/ICASSP.2008.4518793.

Pei-Yun Hsueh and Johanna D. Moore. Combining Multiple Knowledge Sources for Dialogue Segmentation in Multimedia Archives. In *Proceedings of ACL 2007*, 2007. URL https://www.era.lib.ed.ac.uk/handle/1842/4169.

Pei-Yun Hsueh, Johanna D. Moore, and Steve Renals. Automatic Segmentation of Multiparty Dialogue. In *Proceedings of EACL 2006*, 2006.

Shaojun Ji. What do paragraph divisions indicate in narrative texts? *Journal of Pragmatics*, 40(10):1719–1730, October 2008. ISSN 0378-2166. doi: 10.1016/j.pragma.2007.11.010. URL http://www.sciencedirect.com/science/article/pii/S0378216607002159.

Yangfeng Ji and Jacob Eisenstein. Representation Learning for Text-level Discourse Parsing. In *ACL (1)*, pages 13–24, 2014. URL http://www.cc.gatech.edu/grads/y/yji37/papers/ji-acl-2014.pdf.

Ondrej Klejch, Peter Bell, and Steve Renals. Sequence-to-Sequence Models for Punctuated Transcript Combining Lexical and Acoustic Features. In *Proceedings of ICASSP 2017*, New Orleans, USA, 2017. URL http://www.research.ed.ac.uk/portal/files/31745649/icassp_2017_1.pdf.

Alistair Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, University of Edinburgh, Edinburgh, July 1996. URL https://www.era.lib.ed.ac.uk/handle/1842/583.

Alistair Knott and Robert Dale. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62, July 1994. ISSN 0163-853X. doi: 10.1080/01638539409544883. URL http://dx.doi.org/10.1080/01638539409544883.

Greg Kochanski, Esther Grabe, John Coleman, and Burton Rosner. Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118(2):1038–1054, 2005.

J. Kreiman. Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics*, 10(2):163–175, 1982. ISSN 0095-4470(Print).

Catherine Lai, Mireia Farrús, and Johanna Moore. Automatic Paragraph Segmentation with Lexical and Prosodic Features. In *Proceedings of Interspeech 2016*, San Francisco, CA, USA, 2016.

Quoc V. Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In *Proceedings of the 31 st International Conference on Machine Learning*, Beijing, 2014. URL http://arxiv.org/abs/1405.4053.

Ilse Lehiste. Some Phonetic Characteristics of Discourse. *Studia Linguistica*, 36(2):117–130, 1982. ISSN 1467-9582. doi: 10.1111/j.1467-9582.1982.tb00716.x. URL http://onlinelibrary.wiley.com.ezproxy.is.ed.ac.uk/doi/10.1111/j.1467-9582.1982.tb00716.x/abstract.

Gina-Anne Levow. Assessing Prosodic and Text Features for Segmentation of Mandarin Broadcast News. In *Proceedings of SpeechIR'04*, pages 28–32, 2004a.

Gina-Anne Levow. Prosodic cues to discourse segment boundaries in human-computer dialogue. In *Proc. of SIGdial*, 2004b. URL http://www.aclweb.org/old_anthology/W/W04/W04-2318.pdf.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*, 2015. URL https://arxiv.org/abs/1506.01057.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proc. ACL: System Demonstrations*, pages 55–60, 2014.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Iain McGee. The pragmatics of paragraphing English argumentative text. *Journal of Pragmatics*, 68:40–72, July 2014. ISSN 0378-2166. doi: 10.1016/j.pragma.2014.04.002. URL http://www.sciencedirect.com/science/article/pii/S0378216614000770.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2410–2419. JMLR. org, 2017.

Shin'ya Nakajima and James F. Allen. A Study on Prosody and Discourse Structure in Cooperative Dialogues. *Phonetica*, 50(3):197–210, 1993. ISSN 1423-0321, 0031-8388. doi: 10.1159/000261940. URL http://www.karger.com/doi/10.1159/000261940.

Karthik Nandakumar, Yi Chen, Sarat C Dass, and Anil K Jain. Likelihood ratio-based biometric score fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):342–347, 2008.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. SITS: A Hierarchical Nonparametric Model Using Speaker Identity for Topic Segmentation in Multiparty Conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 78–87, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390524.2390536.

J. Niekrasz and J.D. Moore. Unbiased discourse segmentation evaluation. In *2010 IEEE Spoken Language Technology Workshop (SLT)*, pages 43–48, December 2010. doi: 10.1109/SLT.2010.5700820.

John Niekrasz. *Toward Summarization of Communicative Activities in Spoken Conversation*. PhD thesis, University of Edinburgh, 2012.

Aasish Pappu and Amanda Stent. Automatic Formatted Transcripts for Videos. In *Proc. Interspeech*, 2015.

Rebecca J. Passonneau and Diane J. Litman. Intention-based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *Proc. ACL*, pages 148–155, Stroudsburg, PA, USA, 1993a.

Rebecca J. Passonneau and Diane J. Litman. Intention-based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *Proc. ACL*, pages 148–155, 1993b.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Lev Pevzner and Marti A. Hearst. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28(1):19–36, March 2002. ISSN 0891-2017. doi: 10.1162/089120102317341756. URL http://dx.doi.org/10.1162/089120102317341756.

Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum. Unsupervised Topic Modelling for Multi-party Spoken Discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 17–24, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220178. URL http://dx.doi.org/10.3115/1220175.1220178.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 2010. ELRA.

Martin Riedl and Chris Biemann. TopicTiling: A Text Segmentation Algorithm Based on LDA. In *Proc. ACL: Student Research Workshop*, pages 37–42, 2012.

Andrew Rosenberg and Julia Hirschberg. Story Segmentation of Broadcast News in English, Mandarin and Arabic. In *Proc. HLT-NAACL*, pages 125–128, 2006.

Deborah Schiffrin. *Discourse markers*. Cambridge University Press, 1987. ISBN 0-521-35718-7.

Qinfeng Shi, Yasemin Altun, Alex J. Smola, and S. V. N. Vishwanathan. Semi-Markov Models for Sequence Segmentation. In *EMNLP-CoNLL*, pages 640–648, 2007. URL http://www.aclweb.org/old_anthology/D/D07/D07-1.pdf#page=674.

Elizabeth Shriberg. To 'errrr'is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31 (1):153–169, 2001.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1–2):127–154, 2000.

Caroline L. Smith. Topic transitions and durational prosody in reading aloud: production and modeling. *Speech Communication*, 42 (3–4):247–270, April 2004. ISSN 0167-6393. doi: 10.1016/j.specom.2003.09.004. URL http://www.sciencedirect.com/science/article/pii/S0167639303001171.

Noah A. Smith. *Linguistic Structure Prediction*. Morgan & Claypool Publishers, June 2011. ISBN 978-1-60845-406-8.

Caroline Sporleder and Mirella Lapata. Broad Coverage Paragraph Segmentation Across Languages and Domains. *ACM Trans. Speech Lang. Process.*, 3(2):1–35, 2006.

Heather A. Stark. What do paragraph markings do? *Discourse Processes*, 11(3):275–303, July 1988. ISSN 0163-853X. doi: 10.1080/01638538809544704. URL http://dx.doi.org/10.1080/01638538809544704.

S. Strassel. Simple Metadata Annotation Specification V5.0. *Linguistic Data Consortium, Philadelphia*, 2003.

Marc Swerts. Prosodic features at discourse boundaries of different strength. *The Journal of the Acoustical Society of America*, 101(1): 514–521, January 1997. ISSN 0001-4966. doi: 10.1121/1.418114. URL http://scitation.aip.org.ezproxy.is.ed.ac.uk/content/asa/journal/jasa/101/1/10.1121/1.418114.

Marc Swerts and Ronald Geluykens. The Prosody of Information Units in Spontaneous Monologue. *Phonetica*, 50(3):189–196, 1993. ISSN 1423-0321, 0031-8388. doi: 10.1159/000261939. URL http://www.karger.com/doi/10.1159/000261939.

Ryuichi Takanobu, Minlie Huang, Zhongzhou Zhao, Fenglin Li, Haiqing Chen, Xiaoyan Zhu, and Liqiang Nie. A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4403–4410. AAAI Press, 2018.

Ottokar Tilk and Tanel Alumäe. Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. In *Proceedings of Interspeech 2016*, pages 3047–3051, September 2016. doi: 10.21437/Interspeech.2016-1517. URL http://www.isca-speech.org/archive/Interspeech_2016/abstracts/1517.html.

Chiu-Yu Tseng, Zhao-Yu Su, C. Chang, and Chia-hung Tai. Prosodic Fillers and Discourse Markers-Discourse Prosody and Text Prediction. In *Proceedings of TAL 2006*, pages 27–29, 2006. URL http://www.ling.sinica.edu.tw/eip/FILES/publish/2007.4.12.44410341.5339697.pdf.

Emiru Tsunoo, Ondrej Klejch, Peter Bell, and Steve Renals. Hierarchical recurrent neural network for story segmentation using fusion of lexical and acoustic features. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 525–532. IEEE, 2017.

Gökhan Tür, Dilek Hakkani-Tür, Andreas Stolcke, and Elizabeth

Shriberg. Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation. *Computational Linguistics*, 27(1):31–57, March 2001.

Masao Utiyama and Hitoshi Isahara. A Statistical Model for Domain-independent Text Segmentation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 499–506, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics. doi: 10.3115/1073012.1073076. URL http://dx.doi.org/10.3115/1073012.1073076.

Teun A. Van Dijk. Episodes as units of discourse analysis. *Analyzing discourse: Text and talk*, pages 177–195, 1982. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.468.6834&rep=rep1&type=pdf.

Xiaoxuan Wang, Lei Xie, Bin Ma, Engsiong Chng, and Haizhou Li. Phoneme lattice based texttiling towards multilingual story segmentation. In *INTERSPEECH*, pages 1305–1308, 2010. URL http://www.npu-aslp.org/lxie/papers/2010-Interspeech-WangXX-A2-EI-Conf.PDF.

Alexander Yeh. More Accurate Tests for the Statistical Significance of Result Differences. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 947–953, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. doi: 10.3115/992730.992783. URL http://dx.doi.org/10.3115/992730.992783.

Jia Yu, Xiong Xiao, Lei Xie, Eng Siong Chng, and Haizhou Li. A DNN-HMM Approach to Story Segmentation. *Interspeech 2016*, pages 1527–1531, 2016. URL http://www.nwpu-aslp.org/lxie/papers/2016Interspeech-Yujia.pdf.

Margaret Zellers and Brechtje Post. Fundamental frequency and other prosodic cues to topic structure. *Proceedings of IDP 2009*, 2009. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.353.5731&rep=rep1&type=pdf.

L. Zheng, C. C. Leung, L. Xie, B. Ma, and H. Li. Acoustic TextTiling for story segmentation of spoken documents. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5121–5124, March 2012. doi: 10.1109/ICASSP.2012.6289073.

Ji Zhu, Hui Zou, Saharon Rosset, and Trevor Hastie. Multi-class AdaBoost. *Statistics and its Interface*, 2(3):349–360, 2009.